2021 ASEE ANNUAL CONFERENCE
Virtual Meeting | July 26–29, 2021 | Pacific Daylight Time
Paper ID #32916

NSF Data Science Program with Career Support and Connections to Industry

Dr. Carol Shubin, California State University Northridge

Carol Shubin is a professor of mathematics at CSUN and the PI of NSF Data Science Program with Career Support and Connections to Industry. She is interested in partnering with other universities that want to start a data science program. She has been the PI or co-PI in several other STEM educational projects funded by the NSF or NASA and served as a Fulbright Scholar in Rwanda.

CSUN Data Science Program with Career Support and Connections to Industry

Data Science Program with Career Support and Connections to Industry, supported by NSF DUE IUSE, is an interdisciplinary workforce training program that encompasses a summer bootcamp, year-long research projects, biweekly seminars, and career support. Our program has had two cohorts, one in 2019-2020 and the other in 2020-2021. This paper discusses how to design, implement, manage, and assess a data science program for undergraduates.

California State University Northridge (CSUN). CSUN is a federally designated Hispanic Serving Institution (HSI) and Minority Institution (MI). It is among the largest single-campus universities in the United States with enrollments of 39,816 students, drawn largely from the broader Los Angeles County region. Most students (78%) receive some form of financial aid. For these students, college is perceived as a pathway to employment and economic security.

The major goals of our project include:

- increasing the quantity, quality, and diversity of the country's technical workforce by targeting math and computer science students who are interested in a data science career
- offering a solid preparation for careers in data science
- providing support for students to get a job or an internship, or to go to graduate school making participants aware of industry needs
- establishing connections to industry

The students

We recruited a total of 25 seniors or juniors per year with an interest in data science. Most participants majored in computer science or mathematics; however, we also included a few students from business and psychology majors. Applicants were evaluated based on their statement of interest, transcripts, and readiness for the program (i.e. having sufficient math and coding experience). Attention was paid to having a diverse, gender-balanced group. 85% were seniors, 45% were female, and 70% were non-white. We allowed two Deferred Action for Childhood Arrivals (DACA) students to attend although they could not be paid federal money. We also welcomed two students from the NIH Maximizing Access to Research Careers (MARC) program because their lab research was canceled due to COVID-19 restrictions. There was some attrition that we handled by having a waitlist of students who were invited to participate in the program without a stipend. As a few participants dropped out due to time commitments or the difficulty of the program, the waitlist students were more than happy to take their place.

The bootcamp

We offered 6-week intensive summer bootcamps; one was held in the classroom during summer 2019 and other was conducted virtually in summer 2020. The bootcamp ran Monday to Thursday from 9 AM - 4 PM and half-day on Friday. Lectures were followed by lab work. Students were paid $2000 for completing the bootcamp.

The bootcamp covered the six areas of curricula outlined in [1]: (1) data description and curation, (2) mathematical foundations, (3) computational thinking, (4) statistical thinking, (5) data modeling, and (6) communication, reproducibility, and ethics. The recursive data cycle of obtaining, wrangling, curating, managing and processing data, exploring data, defining questions, performing analyses and communicating the results lay at the core of the bootcamp, [2-4].

The topics covered included:

- coding in python and BASH
- coding in python and BASH
- data preprocessing: Pandas
- data exploration and transformation
- feature engineering
- filtering
- wrapper and embedded methods
- machine learning, Scikit-learn, TensorFlow
- data storage: Microsoft SQL Server, MySQL, AWS RDS
- data warehouse server: Microsoft SQL Server Analysis Services (SSAS), Google BigQuery
- cloud computing services: Amazon Web Service (AWS), Microsoft Azure, Google Cloud
- data plotting and visualization: Matplotlib, Basemap, Seaborn, D3 and Google Visualization API
- GIS tools
- Computational environment: Jupyter (IPython) Notebook
- making a Github site

The bootcamp culminated with a choice of week-long projects designed with various levels of difficulty. Most of the mini-projects used datasets from Kaggle, [5], or UCI, [6]. The first year that we ran the project we had trouble coordinating between the four instructors. The second year, we had all teaching materials completed one month in advance of the bootcamp so things ran much smoother.

We preferred students who had a little coding experience, particularly with python, and had taken one introductory statistics class. However, we did admit some students without this preparation.

During our first year, 2019-2020, we did not offer a short introduction to python before we started. Students who were new to coding in python struggled. We realized our error and required students to view a python tutorial, [7], and take three short courses from DataCamp, [8], before the next summerâs program started. In addition, we added a two-week introduction to coding in python in our summer 2020 bootcamp. This was accomplished by subtracting some statistics material. Participants from math and other majors needed this more rigorous introduction to coding. While the computer science majors were somewhat bored, they all said that they still learned a few new things.

One of our Co-PIs created a manual for bootcamp called Getting Started in Machine Learning for Python 3, Scikit-Learn, and Jupyter, [9]. I warmly recommend this book to anyone who wants to teach a hands-on, how-to course in machine learning for undergraduates. While there are mathematical explanations for each method, a beginning student could skip directly to the recipes without any significant loss of continuity throughout the text. The book could be used as a lab manual. Most of the data sets are taken from clean, publicly accessible sources that do not require serious data scrubbing, feature engineering, or other modification.

Year-long projects

During the bootcamp, students were introduced to the team of advisors and their year-long projects. All projects were either supervised by faculty or an engineering lead at a start-up called Nflux AI. Students were paid $500 at the end of each semester. I would strongly advise offering a larger stipend of $1500 - $2000 per semester. Quite a few of our economically challenged students needed outside work which reduced the number of hours that they could spend working on their projects.

Several organizations gave us data: JPL, The Soraya Performing Arts Center, The City of LA, NIH-sponsored Health Equity Research and Education (HERE) Center, NIH-supported BUILD PODER, (Building Infrastructure Leading to Diversity Promoting Opportunities in Diversity, Education, and Research), NFlux, CSUN Institutional Research, and CSUN Autonomy Research Center. It takes time to obtain the data and in some cases the students were required to sign non-disclosure statements. The first year that we ran the program, we underestimated the time it would take to get data and several projects started late due to this oversight. The second year we obtained the data before any of the projects commenced.

In partnership with NASA and JPL, students had the opportunity to work on projects concerned with ecological forecasting, interactive visualization for astrobiological spectroscopy data, and images of Mars. Other projects concerned applications of machine learning to numerical solutions of the kinetic Boltzmann equation, parameter estimation for the spread of COVID-19, traffic sign detection for self-driving cars, quantum modeling projects, and building a 3D-RISM neural network emulator. Some projects came from health, social services, education, and business. They used data science for story-telling. Several of our participants worked for the Health Equity Research and Education (HERE) that studies environmental health factors, neighborhood toxicity, and health disparities in the San Fernando Valley. Other projects leaned heavily on GIS tools; they studied policing in Antelope County and housing and homelessness in Los Angeles. We had a group of students working on projects that had business applications. One project examined Soroya Performing Arts Center subscription data and another optimized revenue for the LA Zoo. There was one project that concerned educational data that asked questions about student performance at CSUN based on high school records, time-to-degree with or without change of majors, and grading differences based on instructional rank. It was good to have a wide range of problems as students had very different interests and technical strengths. Several students participated in more than one project.

Career seminar and linkages to industry

The data science program participants attended a biweekly seminar series during the academic year. The seminar focused on career preparation workshops and guest speakers gave information about their industry, their interview process, and criteria for employee selection. Students created a LinkedIn page, their resumes were critiqued, and engaged in mock interviews. Students were sent information about current job opportunities and internships regularly. They are required to apply for three internships. It has not been easy to establish close connections with employers; however, we have made some headway. Our seminar featured speakers from Google, Air Force Research Laboratory (Edwards and Kirtland), Wells Fargo, Amgen, Zest AI, NFlux AI, HRL, XYPRO, Equifax, JPL, Farmerâs Insurance, PennyMac, Arete, Merck, and CACI International.

Our program is partly concerned with developing linkages to industry and helping students get internships and careers in data science. During Summer 2020, three students had internships at NFlux AI that were conducted remotely. Interns helped to develop technology that enhances and augments human decision-making capabilities. The goals were to solve an assortment of control task problems using a myriad of reinforcement learning methods while creating a scalable system. I received very favorable updates on those students' progress. Other participants had internships at Electronic Warfare Integrated Laboratories Naval Air Warfare Center Weapons Division (NAWCWD), JPL, Microsoft, NFlux AI, Air Force Research Laboratory, NASA Ames in the Intelligent Systems Division, GM, PennyMac, the US Navy, CSUN Autonomy Research Laboratory, the Family Empowerment Center, and the HERE Center.

Institutionalization

Before our program started, data mining and machine learning were offered as Masterâs degree level courses in Computer Science. Few non-computer science students took those classes. However, two of our Co-PIs spearheaded the effort to create a data science minor for undergraduates that commenced this year, [10]. This minor institutionalized two of our goals: offering preparation for careers in data science and increasing the quantity, quality, and diversity of the workforce.

External evaluation

Our external evaluation consisted of a pre- and post-survey that asked questions of the studentsâ background knowledge in python and data science as well as some more personal questions about their activities and ability to work independently. The post-survey also included questions about what the participants liked and did not like about the program. Overall, students expressed âloveâ and âenjoymentâ for the program. Despite the COVID-19 pandemic, many were grateful for the opportunity and they were all happy to participate in something over the summer.

Management

As the PI, it was my job to recruit and keep track of the students, TAs, and advisors. Advisors were contacted every two months for a brief progress report. Students posted written reports a quarter and three-fourths of the way through and gave presentations mid-way through and at the end of the academic year.

I would like to pass on some lessons learned regarding working with a group of co-PIs who do not know each other very well and who come from different departments. Do not underestimate the conflicts that can arise. It took a year to work out all our management problems that stemmed from not sufficiently thinking out our co-PI compensation that was not equally divided. Some co-PIs received more reassigned time for working on the project than others. Here the problem stemmed from the very hurried manner in which the grant was written. Also, faculty were paid one-month summer salary. Different faculty have different salaries so it seemed unfair to those who received less than they should be paid less for the same (or more) amount of work. This could have been easily addressed by paying everyone a flat $10,000. Furthermore, some Co-PIs were very good about preparing their course materials ahead of time. Others were not and the first year was not very well-coordinated. The second summer ran much smoother.

Adjustments due to COVID-19

CSUN, like most other universities, switched to a virtual environment for all instruction in March 2020. All of our year-long projects were also conducted virtually. Final presentations at the end of May 2020 showed that many of the projects experienced some disruption. Furthermore, most of our plans for widespread dissemination about our program were also disrupted by COVID-19.

Covid-19 directly affected several students who became sick or had relatives die. While several of our graduating seniors found jobs, quite a few of the students found the job market in 2020 very challenging. Undoubtedly, the lockdown and economic disruption adversely affected our students' employment opportunities. Many students decided to apply for MS programs in either math or computer science at CSUN because there were no jobs.

The Summer 2021 bootcamp was held virtually and I feel that it went very well. By then the faculty were very competent teaching over ZOOM and the new cohort of students had also adjusted to the situation. We selected four of our most successful graduates from Cohort 1 to be TAs for Cohort 2. We made extensive use of breakout rooms so students did get to interact with each other during labs. Participants reported a very high level of satisfaction with the instruction and were effusive in their praise of the TAs. All the mini-project presentations went well.

Students began working remotely on their year-long projects at the end of August. The mid-term presentations were held on January, 2021 and it appeared that all students were making very good progress. Advisors meet with students weekly over ZOOM. While everyone is looking forward to returning to our traditional format, we have all adjusted and I really believe that this year's program is running as well as it would under traditional circumstances.

References

[1] ”Envisioning the Data Science Discipline: The Undergraduate Perspective: Interim Report” National Academies Press: OpenBook, https://www.nap.edu/read/24886/

[2] Berman, F., Rutenbar, R., Hailpern, B., Christensen, H., Davidson, S., Estrin, D., â Szalay, A. S. Realizing the potential of data science. Communications of the ACM, 61(4), 67-72, 2018. https://doi.org/10.1145/3188721

[3] De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., â Ye, P. Curriculum Guidelines for Undergraduate Programs in Data Science. Annual Review of Statistics and Its Application, 4(1), 15 - 30, 2017. https://doi.org/10.1146/annurev-statistics-060116-053930

[4] Donoho, D. 50 Years of Data Science. Journal of Computational and Graphical Statistics, 26(4), 745 - 766, 2017. https://doi.org/10.1080/10618600.2017.1384734

[5] Your Machine Learning and Data Science Community. Kaggle. https://www.kaggle.com/

[6] UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/index.php

[7] YouTube. (2018, July 11). Learn Python - Full Course for Beginners [Tutorial], Giraffe Academy, YouTube. https://www.youtube.com/watch?v=rfscVS0vtbw

[8] DataCamp Courses: Introduction to Python, Introduction to Data Science, and Statistical Thinking in Python, https://www.datacamp.com

[9] Romero, B. (aka Bruce E. Shapiro). Getting Started in Machine Learning for Python 3, Scikit-Learn, and Jupyter, Sherwood Forest Books, Los Angeles, 2020.

[10] Minor in Data Science Minor, Computer Science, CSUN University Catalog, California State University, 2020. https://catalog.csun.edu/academics/comp/programs/minor-data-science/